bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Deep Learning from Phylogenies for Diversification Analyses

AUTHORS

Lambert Sophia[1†*], Voznica Jakub[2,3,4†*], Morlon Hélène[1]

AFFILIATIONS

[1]Institut de Biologie de l'École Normale Supérieure, École Normale Supérieure, CNRS, INSERM, Université Paris Sciences et Lettres, 75005 Paris, FRANCE;

[2]Unité de Bioinformatique Évolutive - Département Biologie computationnelle, Institut Pasteur, Paris, FRANCE;

[3]Unité de Biologie Computationnelle, USR 3756 CNRS, Paris, FRANCE;

[4]Université Paris Cité, Paris, FRANCE;

† These authors contributed equally to this work

** corresponding authors: slambert@bio.ens.psl.eu and voznica.jakub@gmail.com*

ABSTRACT

Birth-death models are widely used in combination with species phylogenies to study past diversification dynamics. Current inference approaches typically rely on likelihood-based methods. These methods are not generalizable, as a new likelihood formula must be established each time a new model is proposed; for some models such formula is not even tractable. Deep learning can bring solutions in such situations, as deep neural networks can be trained to learn the relation between simulations and parameter values as a regression problem. In this paper, we

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

adapt a recently developed deep learning method from pathogen phylodynamics to the case of diversification inference, and we extend its applicability to the case of the inference of state-dependent diversification models from phylogenies associated with trait data. We demonstrate the accuracy and time efficiency of the approach for the time constant homogeneous birth-death model and the Binary-State Speciation and Extinction model. Finally, we illustrate the use of the proposed inference machinery by reanalyzing a phylogeny of primates and their associated ecological role as seed dispersers. Deep learning inference provides at least the same accuracy as likelihood-based inference while being faster by several orders of magnitude, offering a promising new inference approach for deployment of future models in the field.

KEYWORDS: convolutional neural networks, birth-death models, deep learning, phylogeny representation, diversification, macroevolution.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

Phylogenetic approaches for studying species origination and extinction dynamics over deep time rely on the statistical adjustment of stochastic birth-death models (Kendall 1948) to dated phylogenetic trees representing the evolutionary relatedness of species and the dating of their divergence times (Stadler 2013; Morlon 2014; Harmon 2019). An increasing amount of such phylogenetic data have become available, and has been accompanied by complexification of diversification models. These models include homogeneous rate models where speciation and extinction rates are identical across lineages at any given time, and range from simple time-constant (Nee et al. 1994) to time-dependent (Morlon et al. 2011; Stadler 2011; May et al. 2016), environment-dependent (Condamine et al. 2013), and diversity-dependent (Etienne et al. 2012) models. Diversification models also include heterogeneous rate models (Alfaro et al. 2009; Morlon et al. 2011; Rabosky 2014; Höhna et al. 2019; Maliet et al. 2019; Barido-Sottani et al. 2020; Laudanno et al. 2020), with the class of State-dependent Speciation and Extinction (SSE) models that links rate heterogeneity to specific characteristics of the species (Maddison et al. 2007; FitzJohn 2010; Goldberg et al. 2011; Fitzjohn 2012; Beaulieu and O'Meara 2016; Herrera-Alsina et al. 2019; Vasconcelos et al. 2022).

The parameters of interest of these models – mainly speciation and extinction rates – are traditionally inferred using likelihood-based techniques – Maximum likelihood or Bayesian inference. Maximum likelihood consists in finding the parameters that maximize the probability of observing the data (here the phylogeny, or the phylogeny and associated trait data in the case of SSE models); Bayesian inference uses this probability along with *a priori* information on the parameters of interest to explore the *posterior* probability distribution of the parameters knowing the data. Numerous studies have used these inference approaches to estimate speciation and

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

extinction rates over geological times and across the Tree of Life, to investigate the processes modulating these diversification dynamics (Stadler 2011; Etienne et al. 2012; Pyron and Wiens 2013; Höhna 2014; Rolland et al. 2014; Gubry-Rangin et al. 2015; Rabosky et al. 2018; Condamine et al. 2019; Stone and Wolfe 2021). While powerful, likelihood-based inference techniques are limited by potential intractability issues for complex diversification models, and by computational cost on increasingly large phylogenetic data (Hinchliff et al. 2015). Indeed, complex models do not always have a closed-form solution and/or the likelihood cannot always be evaluated in a reasonable amount of time. This renders the application of likelihood-based methods to complex models and species-rich groups (*e.g.*, insects, micro-eukaryotes and prokaryotes) difficult. As a result, there are several models of diversification in the literature which behavior has been studied with simulations but that lack a proper inference machinery (McPeek 2008; Aristide and Morlon 2019; Hagen et al. 2021). Methods based on Expectation Maximization (EM) algorithms (Dempster et al. 1977; Richter et al. 2020), data augmentation (Maliet and Morlon 2022), or composite likelihoods (Lindsay 1988; Varin et al. 2021)) can overcome some of these limitations (Raynal 2019), yet they still rely on likelihood formulae.

An alternative is the use of likelihood-free inference techniques, such as Approximate Bayesian Computation (ABC (Beaumont et al. 2002; Marin et al. 2012; Sisson et al. 2018)). In its most basic form, ABC relies on generating artificial data by simulating the process of interest along a given parameter range and compressing the data by computing summary statistics on these simulations to enable the comparison between simulated and observed summary statistics. This comparison is done by computing a distance and evaluating if this distance is sufficiently small to accept the simulated data using a chosen tolerance threshold (rejection-based approach). ABC has been useful to fit complex models in various fields, including phylogenetic

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

diversification analyses (Bokma 2010; Janzen et al. 2015; Janzen and Etienne 2016). However, there are several limitations of ABC, including its reliance on the choice of summary statistics that should resume the information contained in data in a low number of metrics, the choice of the distance metric to compare the observed and simulated data and the choice of the tolerance threshold. While it is possible to evaluate and minimize the influence of these choices on parameter inference (Sisson et al. 2007; Beaumont et al. 2009; Blum and François 2010; Del Moral et al. 2012; Blum et al. 2013; Prangle 2017), an adjustment needs to be performed each time a new model is developed. In particular, new summary statistics must be designed to convey information relative to the problem at hand.

Deep learning offers an alternative likelihood-free inference technique. Deep learning (Goodfellow et al. 2016) is a subfield of machine learning where highly flexible statistical learning functions based on neural networks (NNs) are used to learn regression (such as parameter estimation) or classification (such as model selection) problems. The term ‘deep’ is conventionally associated with a neural network that takes raw data as input values and extracts patterns from this low level representation thus creating its own ‘summary statistics’ or high-level features, without the need of designing those. This definition is the one used in the present article. Applying a regression task with deep learning (learning model parameter values from simulations) is increasingly used in several fields including population genetics (Sheehan and Song 2016; Sanchez et al. 2020; Avecilla et al. 2022), phylogenetic reconstruction (Nesterenko et al. 2022), macroecology (Andermann et al. 2022) and physiology (Kroll et al. 2021). In macroevolution, an early progress of using machine learning (Bokma 2006) consisted in training an artificial neural network on the axes of a principal component analysis of phylogenetic branching times to infer speciation and extinction rates. A similar framework was then used for

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the inference of rates of phenotypic evolution (Bokma 2010). While promising, this approach was not developed further by the community.

A step forward was recently taken by Voznica et al. (2022), who developed a deep learning approach for the statistical inference of birth-death models from phylogenies in the context of pathogen phylodynamics. The authors developed a tree representation, the Complete Bijective Ladderized (CBLV) tree representation, which applies to non-ultrametric trees representing the evolutionary relationship between pathogen sequences sampled at different dates. The CBLV representation proved to be efficient for model selection and inference of transmission dynamics when combined with Convolutional Neural Networks (CNN) (LeCun et al. 1998), where it yielded accuracy at least comparable to gold-standard Bayesian approaches. Voznica et al. (2022) also combined an extensive set of Summary Statistics with Feed-Forward Neural Networks (FFNN-SS) that yielded similar results. To our knowledge, comparable attempts to use deep learning to infer diversification dynamics from species phylogenies do not exist.

Here, we adapt the approach of Voznica et al. (2022) to birth-death diversification models used to infer diversification dynamics from phylogenies of extant species (and also potentially associated trait data). The phylogenetic data are different from the one in Voznica et al. (2022). First, the evolutionary trees are different as, if no data from fossils is included, the species are all sampled at present time and the reconstructed phylogenies are thus ultrametric, that is the tips are all at the same distance to the root. Second, we allow for possibility to include trait data associated to the tips. We begin by adapting the CBLV tree representation from Voznica et al. (2022) to ultrametric phylogenies and to ultrametric phylogenies with tip state data, in the form of ‘Complete Diversity-reordered Vector’ (CDV). We then assess the

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

127 performance of deep learning inference in comparison to maximum likelihood estimation (MLE)
128 by using the simple homogeneous time-constant birth-death (BD) model (Nee et al. 1994), for
129 which a closed-form expression of the likelihood exists, and the binary-state speciation and
130 extinction (BiSSE) model, for which the likelihood is approximated by solving Ordinary
131 Differential Equations (Maddison et al. 2007). Finally, we illustrate the approach by applying our
132 trained neural network for BiSSE to an empirical phylogeny of 273 primates (Fabre et al. 2009)
133 and their associated interaction type (mutualistic or antagonistic) with plants (Gómez and Verdú
134 2012).
135

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

METHODS

Our main goal is to develop a compact and exhaustive representation of the raw data into a matrix (the CDV), and to test the performance of a CNN combined with this representation for parameter inference (thereafter referred to as the CNN-CDV approach). We train the neural networks with data simulated under the birth-death diversification processes. For the time-homogeneous birth-death model (BD), we compare the performance of CNN-CDV to FFNN combined with a series of summary statistics (FFNN-SS), FFNN combined with the CDV representation (FFNN-CDV), and the maximum likelihood estimation (MLE) approach. For the BiSSE model, we compare the performance of CNN-CDV to FFNN-CDV and MLE. Codes used to perform the simulations, encode the phylogenetic data into the CDV representation, train the neural networks, and use the trained networks on simulated or empirical data for parameter inference, are available on GitHub (https://github.com/JakubVoz/deeptimelearning).

TREE REPRESENTATIONS

Full Tree Representation: Compact Diversity-reordered Vector (CDV)

We adapted the Compact Bijective Ladderized Vector (CBLV) representation used in Voznica et al. (2022), originally intended for non-ultrametric trees, to ultrametric trees and to ultrametric trees with information on tip states. The encoding proceeds in the following steps (**Fig. 1)** : 1] internal node reordering, 2] inorder tree traversal (Cormen 2009) and creation of vector representation, 3] completion to the maximum tree-size in simulations, 4] addition of tree height and sampling probability to the vector representation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

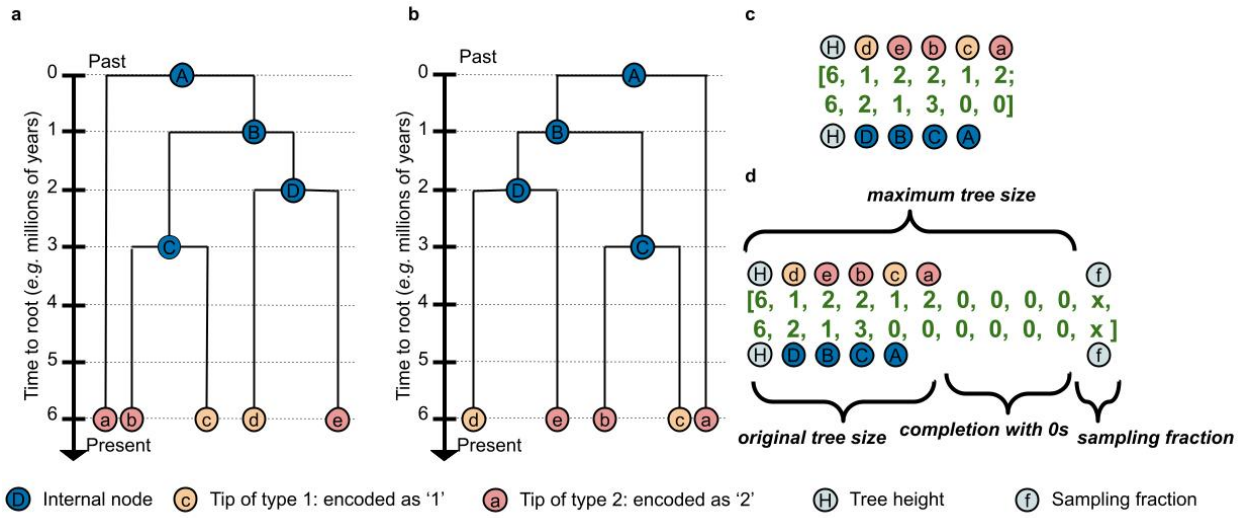

Figure 1: Representation of ultrametric trees with tip state data.

Illustration of the encoding algorithm for 'Compact Diversity-ordered Vector' (CDV) on a tree with 5 tips. ***(a)*** *Ultrametric tree with tip state information (here there are two possible states for each tip, either state 1 or state 2, represented in orange and pink respectively).* ***(b)*** *The tree is reordered following a diversity criterion: for each internal node, the sum of the branch lengths of the descending tree is computed and the internal node with higher sum is rotated on the left.* ***(c)*** *We then create a 2-rows matrix filled by visiting the tree according to a tree inorder traversal algorithm, with tips represented on the top row and internal nodes on the bottom row. Tips are assigned their encoded trait state ('1' for state 1 and '2' for state 2) and internal nodes their distance to the root. We add a first column with tree height.* ***(d)*** *Finally, we complete the matrix with zeroes, so that its size is the one of the largest simulated tree and we add a column with the value of the sampling fraction. To obtain the representation of an ultrametric tree without tip state data, we do not include the first row on tip state information.*

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Before encoding, we rescaled the trees to unit average branch length and the rate parameters (*e.g.* diversification rate(s), turnover rate(s) *etc.*) accordingly. The trained neural networks should thus apply to trees with any time scale. The criterion used in Voznica et al. (2022) for tree reordering was based on the ladderization, where each internal node is rotated so that the branch supporting the most recent tip is on the left. As this cannot apply to ultrametric trees where all tips are sampled at the same time, we used a diversity criterion: for each internal node we compute the sum of the branch lengths of the descending tree (i.e. phylogenetic diversity) and the branch with highest sum is shifted to the left. Next, we perform a tree inorder traversal: the reordered phylogeny is traversed by recursively starting with the left subtree and for each visited internal node, its distance to the root is added to a vector (Cormen 2009). For trees with tip data, we create a second vector with information on the tip data, while visiting tips during the same traversal. For binary tip data (as obtained by simulating BiSSE for example), we use the values 1 and 2 to distinguish the two states. These two vectors are then combined into a matrix. We then add a first column with the value of the tree height. We name this representation a Compact Diversity-ordered Vector (CDV). This representation could be easily extended to account for information on non-binary traits, for example using one hot encoding, which consists in encoding a qualitative variable of x states into x rows with 1 representing the presence of state and 0 its absence. Similarly, multiple traits, including quantitative ones, can be encoded by stacking additional rows to the matrix. Furthermore, we can imagine treating missing trait data by assigning a value of -1 in the CDV representation when the information is lacking.

The CDV representation is bijective (under mild assumptions, for example the absence of concomitant branching events), in the sense that we can unambiguously reconstruct any given tree from its representation, and compact: (x+1)**n* entries for a tree with *n* tips (n-1 values for

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

internal nodes and 1 on tree height) and x informational values on tips (x=0 in the absence of tip state information and x=1 for information on a single trait). The created vector (or matrix when tip state information is available) is completed with zeroes to obtain a representation of the same size as the largest phylogeny in the simulations. In order to account for potential missing extant species in the phylogeny, we add a last column with the value of the sampling fraction, computed as the ratio of the number of species represented in the phylogeny divided by the total number of extant species.

In order to assess whether the CDV representation with tip data allows to properly capture tip information, we compared results obtained with this representation to those obtained with a less informative representation (coined CDV-less) where we add only two values on tip type counts (*i.e.* the number of tips in each state) at the end of the row summarizing internal nodes.

Summary statistics representation

We used a set of 97 summary statistics (SS) representing trees (without associated trait data as we do not implement the FFNN-SS for the BiSSE model). The summary statistics were mainly based on those published in Saulnier et al. (2017) and in Voznica et al. (2022) (see the original papers for details) and they comprise:

- 8 summary statistics on tree topology
- 25 summary statistics on branch lengths
- 49 summary statistics on the Lineage-Through-Time (LTT) plot
- 14 summary statistics on consecutive internal branches
- 1 summary statistics on number of tips

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We modified several statistics so that they apply to ultrametric trees. Instead of minimal and maximal tree height (the time of first and last sampled tips in non-ultrametric trees), we use the crown age. In Saulnier et al. (2017), there are several summary statistics defined on the LTT plot which consider the maximum number of the living lineages, the time of its occurrence and the slopes of the curves before and after this time. In ultrametric trees, the maximum number of the living lineages always appears at present (the moment of sampling tips) and thus such division of LTT plot is not possible. Instead, we divide the LTT plot into three equal parts and we measure the slope for each one, together with the ratios between the first and the second slope and between the second and the third slope. The computing time of these statistics grows linearly with tree size. We added the sampling fraction to these 97 measures, thus resulting in a vector of 98 scalars.

Finally, we reduced and centered the SS by subtracting the mean and scaling to unit variance, using the standard scaler from the scikit-learn package (Pedregosa et al. 2011) fitted to the training set.

NEURAL NETWORKS: ARCHITECTURE AND TRAINING

A NN is organized in neural layers that in turn are organized in neurons (or 'units'). In supervised learning, a NN can be trained to minimize the difference between an expected output (or target) and the predicted one, the measure of the difference being called a 'loss function'. Here we used the mean absolute error as the loss function. A NN contains an input layer (by which the numerical values representing the data are passed to the following layer) and an output layer (in our case outputting parameter values), potentially separated by hidden layers. If there is at least one hidden layer, we talk about deep neural networks.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Feed-Forward Neural Networks (FFNNs) are one of the most basic forms of deep NNs. They are fully connected: for each neural layer, all neurons are connected to the neurons of the previous layer. The connections are characterized by trained bias and weights *i.e*. real values by which individual inputs of a given neuron are multiplied, as well as an activation function by which the summed input (input values multiplied by weights to which a bias value is added) is transformed. Here, we used the exponential linear function as activation function (Clevert et al. 2015). FFNNs typically work well on structured data, such as summary statistics, where the same summary statistics are at exactly the same entry in the input vector. On unstructured data (such as images or CDV) however, the number of parameters to train increases quadratically with the size of the input, and the information is scattered along the input vector, making FFNNs potentially less efficient.

Convolutional Neural Networks (CNNs) (LeCun et al. 1998) contain a convolutional part and a fully connected one. The convolutional part consists of convolutional and pooling layers and outputs a vector used as input of the fully connected part. The convolutional part aims at learning and extracting repeated patterns in the input, that are then combined for prediction in the fully connected part. Convolutional layers transform their input with several convolutional operations, each one specified by a kernel (or 'filter' or 'feature detector') whose parameters are trained. Each kernel transforms subparts or patches of the input by applying the convolutional operation and outputs a single value for each patch. We set the stride (by how much the kernel moves on the input when traversing it) to 1. A convolutional layer is specified by the number of kernels and their size (the size of their input), and outputs a 'feature map' (intermediate representation of transformed input). Pooling layers transform the resulting feature maps into smaller ones by taking the maximum or average of values subsampled in the map with a given

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

window. CNNs typically work well with raw, low-level (vector and matrix) data such as images, videos or time-series recordings, by learning and extracting repeated patterns or 'features' through trained convolution functions and building from them their own high-level features (such as a set of summary statistics). They do not need prespecified feature input such as summary statistics. Each kernel learns and extracts one pattern in the data that can appear anywhere in the representation. In comparison to Voznica et al. (2022), increased kernel size in the first convolutional layer performed slightly better (see below; data not shown).

The training of a given network consists in iteratively changing the parameters of the NN (*e.g.* bias, weights) that minimize the loss function. This is performed by an optimization algorithm for stochastic gradient descent. Here, we used the Adam optimizer (Kingma and Ba 2015). The networks were trained on simulated data, the targets being the parameters of the diversification models (here BD and BiSSE).

Several training 'tricks' were developed for efficient and robust training in practice. We used a training set of 990.000 simulations split into subsets called batches, during the training. After measuring the loss on the whole batch, the trained parameters are updated with the optimizer to minimize the loss. Splitting the training set into batches of simulations enables to update the trained values more robustly (and moving into 'right direction' with respect to the minimal error). We set the batch size to 8,000 simulations. When the network parameters were updated on the whole training set (we talk about an 'epoch'), the training starts again passing through the whole training set.

To prevent overfitting, we used a dropout of 0.5, which consists in shutting down randomly half of the neurons in the network during the training phase (Srivastava et al. 2014). We also used a technic called early stopping (Bengio 2012): at the end of each epoch, the loss is

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

computed on a validation set (here, we used a validation set of 10.000 simulations), and the training is stopped when the loss on the validation set starts to increase. Typically, during the training of our networks, there will be hundreds of epochs before the training is stopped. The test set then enables to measure the true accuracy of the network. In our case it consisted of 500 simulations for the BD model and 10,000 simulations for BiSSE.

For the BD model, we used a CNN architecture on the CDV representation (referred to as CNN-CDV), a FFNN architecture on the CDV representation (FFNN-CDV), and a FFNN architecture on the summary statistics representation (FFNN-SS). For the BiSSE model, we used only the CNN-CDV and the FFNN-CDV, as applying the FFNN-SS would require deploying a new set of summary statistics accounting for tip data. We used the same FFNN architecture as in Voznica et al. (2022) in the case of both the FFNN-CDV and the FFNN-SS (see the original paper for details). The CNN architecture differed only by the size of the kernels (i.e. the size of their inputs) in the first layer; we used layers of size 5*2 and 5 for respectively the CDV and CDV-less as it performed slightly better than kernels of size 3*2 (respectively 3, data not shown). Also, unlike in Voznica et al. (2022), the function of the output layer was set to the exponential linear function (Clevert et al. 2015).

We implemented the NNs in Python 3.6 using the Tensorflow 1.5.0 (Abadi et al. 2016), Keras 2.2.4 (Chollet 2015) and scikit-learn 0.19.1 (Pedregosa et al. 2011) libraries.

MACROEVOLUTIONARY MODELS AND SIMULATIONS

We assessed the performance of the NNs with two widely used models, the simple homogeneous, time constant birth-death (BD) model, and the binary-state speciation and extinction (BiSSE) model.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Birth-Death (BD) Model

The BD model has a closed-form expression of its likelihood, which allows us to compare the performance of the Deep Learning and MLE approach in a "best case" scenario for MLE. Here, when referring to BD, we imply the homogeneous time-constant birth-death-sampling model (Yang and Rannala 1997; Stadler 2009), as we consider the possibility that some extant species are not represented in the phylogenies. In this model, new species originate with a constant speciation rate λ and go extinct with a constant extinction rate μ, typically expressed in number of events/lineage/Myr. At present each extant species is sampled with probability f (Bernoulli sampling scheme (Stadler 2009)). We assume f to be fixed, in which case λ and μ are identifiable.

We parametrized our simulations with the turnover ($\varepsilon = \mu / \lambda$) and speciation rate (**Table 1**); the parameter values were sampled uniformly at random within parameter boundaries with standard Latin-hypercube sampling (McKay et al. 1979) using the Python PyDOE package. We performed the simulations with our own simulator, using a Gillespie algorithm (Gillespie 1977). Each simulation started with one lineage and ended when the number of living species reached $\frac{s}{f}$, where s is the number of tips in the sampled phylogeny. We then sampled s species. We thus conditioned the simulations on the number of tips. We trained the NNs to learn λ and ε.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1: Parameterization of the constant-rate birth-death model with incomplete sampling.

Parameters	Symbols	Ranges
Turnover rate	**ε**	**U(0.01,1)**
Speciation rate	**λ**	**U(0.01,0.5)**
Tree size	s	U(200, 500)
Sampling fraction	f	U(0.01,1)

For each parameter, we display its symbol and the range of values used in the simulations by sampling from the uniform distribution (indicated with U) in training, validation and testing sets. Parameters indicated in bold are those that are estimated during the inference.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The Binary State Speciation and Extinction (BiSSE) Model

The BiSSE model is the simplest state-dependent birth-death model: species are characterized by a binary state (1 or 2) which can influence their constant speciation (λ_1 and λ_2) and extinction rates (μ_1 and μ_2). Species can also transition anagenetically from one state to the other (with rates q_{12} and q_{21}). As for the BD model, we can add to this diversification process a Bernoulli sampling scheme at present, which allows analyzing trees with missing extant species. We consider here a simple version of BiSSE with symmetrical transition rates $q=q_{12}=q_{21}$ as well as turnover rate and sampling probabilities at present (ε and f, respectively) shared across species irrespective of their state.

The BiSSE model allows us to illustrate the utility, and test the validity of the CDV representation with tip data. The likelihood of this model can be computed by solving Ordinary Differential Equations (ODEs), and recent efforts have been made to provide an efficient maximum-likelihood inference machinery for this model on large phylogenies (Louca and Pennell 2020).

We parameterized our simulations with the speciation rate associated to state 1 (λ_1), the turnover rate (ε) and the ratios of λ_2 and q_{12} relative to λ_1. We sampled these parameters within a biologically realistic parameter space, given the literature on empirically inferred parameters (e.g. (Villarreal and Renner 2013; Williams et al. 2014; Gamisch 2016)) (**Table 2**). The parameter subspace was covered with standard Latin-hypercube sampling (McKay et al. 1979) using the Python PyDOE package. The simulations were performed using the R package castor 1.6.6 (Louca et al. 2018) and were conditioned on number of tips. We trained the NNs to learn λ_1, λ_2, q_{12} and ε.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 2: Parameterization of the Binary State Speciation and Extinction model with incomplete sampling.

Parameters	Symbols	Range
Turnover rate	**ε**	**U(0,1)**
Speciation rate 1	**λ_1**	**U(0.01,1)**
Speciation rate 2 and its ratio to λ_1	**λ_2**, $r_{\lambda 2}$	**[10e-3,1]** U(0.1,1)
Transition rate and its ratio to λ_1	**q_{12}=q_{21}**, r_q	**[10e-4,0.1]** U(0.01,0.1)
Tree size	s	U(200, 500)
Sampling fraction	f	U(0.01,1)

For each parameter, we display its symbol and the range of values used in the simulations by sampling from the uniform distribution (indicated with U) in training, validation and testing sets. Parameters indicated in bold are those that are estimated during inference. λ_2 is parameterized with respect to λ_1, being at 10% to 100% of its value. The transition rates are also parameterized with respect to λ_1, being at 1% to 10% of its value. The corresponding ranges of values are indicated in brackets.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MAXIMUM LIKELIHOOD ESTIMATION

We compared the NN approaches with MLE. For the constant-rate BD, we used a MLE based on the Nelder–Mead optimization algorithm encoded within our custom function fitMLE_bdRho available at https://github.com/sophia-lambert/UDivEvo/tree/master/R. The function encodes a likelihood formula conditioned on the age of the phylogeny (here $t_0 = t_{crown}$) under a Bernoulli sampling scheme, and parametrized to infer the net diversification rate ($r = \lambda - \mu$) and the turnover rate (ε) as it is easier to maximize the likelihood in this reparametrized likelihood landscape. For the BiSSE model, we used the MLE deployed under the R packages diversitree 0.9-3 (Fitzjohn 2012) and castor 1.6.6 (Louca and Doebeli 2018), both conditioned on the crown age of the phylogeny, under a Bernoulli sampling scheme and parametrized to infer λ_1, λ_2, μ_1, μ_2, and $q=q_{12}=q_{21}$. Diversitree is the traditional package used for fitting BiSSE; castor was developed more recently, and implements a faster algorithm for computing the likelihood of SSE models on large phylogenies (Louca and Doebeli 2018).

PERFORMANCE ASSESSMENT

Accuracy of Parameter Estimation

To assess the accuracy of parameter estimation, we used 500 simulated test trees for the simple BD model and 10 000 for BiSSE. 17 BiSSE simulations for which castor and/or diversitree outputted an error message or no estimated values (15 for castor, 3 for diversitree with one simulation in common) were excluded from these analyses, resulting in 9.983 test trees.

To avoid over-penalizing the MLE approaches that, contrary to the NNs, do not have constrained parameter ranges, we added similar constraints to the MLE estimates. Indeed, to the exception of λ for constant-rate BD and λ_1 for BiSSE for which NNs can predict values outside

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of the parameter values initially covered by the simulations due to tree rescaling, the other parameter estimates (ε, λ_2 and $q=q_{12}=q_{21}$) are constrained by the parameter range used for ε, $r_{\lambda 2}$ and r_q in our simulations. We imposed similar constraints to the MLE estimates to prevent overpenalizing MLE in accuracy comparison. For example, if the MLE for ε is above 1, we set it to 1, and if the MLE for λ_2 is lower than $0.1*\lambda_1$, we set it to $0.1*\lambda_1$ (0.1 is the minimum value for $r_{\lambda 2}$ used in our simulations).

To evaluate the distribution of the bias among the test set, we calculated the bias between the true (simulated, or 'target') parameter values and the predicted values per predictions as follows:

- *Bias* $B_i = \text{predicted}_i - \text{target}_i$

We computed three measures of error between the true parameter values and the predicted values on the test set:

- *Mean absolute error* $MAE = 1/n * \sum_i^n abs(\text{predicted}_i - \text{target}_i)$
- Mean relative absolute error $MRE = 1/n * \sum_i^n abs(\text{predicted}_i - \text{target}_i)/\text{target}_i$
- *Mean bias* $MB = 1/n * \sum_i^n (\text{predicted}_i - \text{target}_i)$

We also report the Pearson correlation coefficient between simulated and predicted values, computed with the R package 'stats' ('cor.test' function) version '4.2.1'.

Finally, we assessed the influence of tree size on parameter estimation accuracy.

Time Efficiency

We compared the average time of estimation between CNN-CDV and MLE for BiSSE. For the CNN-CDV approach, we reported the average CPU time of encoding a tree (averaged over 1,000,000 trees). The estimation on itself is negligible with respect to the time of encoding. For

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

MLE BiSSE estimation with the castor and diversitree packages, we reported the average CPU time (average over 10.000 test trees, out of which 18 resulted in an error).

EMPIRICAL ILLUSTRATION

Primates can have an antagonistic interaction with plants through herbivory, or a mutualistic one through frugivory and seed dispersal (Gómez and Verdú 2012). To illustrate the application of our deep learning approach, we reanalyzed the dataset of Gómez and Verdú (2012), that categorizes primate species according to the nature of their interaction with plants (state 1 for a mutualistic interaction, and 2 for an antagonistic one), using the primates phylogeny of Fabre et al. (2009). We started by pruning taxa without information on interaction type (13/273) and rescaled the phylogeny as previously described. Next, we performed some sanity checks to verify that the empirical data fell within the space covered by our BiSSE simulations; if it does not, this means that the model and/or the range of parameters used in the simulations is not well adapted to the empirical data, in which case application of the trained neural network to the data might output meaningless results. For these analyses, we used the set of phylogenies that we simulated under the BiSSE model to produce our test set. First, we checked that each of the summary statistics values calculated on the empirical data fell within the range spanned by the summary statistics values calculated on the simulated phylogenies. Then, we performed a principal component analysis (R package 'FactoMineR', function 'PCA') on the set of summary statistics described above (see Methods) with the addition of four simple summary statistics adapted to the inclusion of tips data: the number of tips in each state (1 or 2) and the phylogenetic diversity of each state (R package 'picante', function 'pd'). Finally, we transformed the data into its CDV representation and fed it to the trained CNN-CDV network on the BiSSE

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model. We used a sampling fraction of 0.68, computed using a global diversity of 381 stable species complexes for primates, following Gómez and Verdú (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESULTS

PERFORMANCE ASSESSMENT

The comparison of parameter estimates obtained using deep NNs versus MLE for the BD model shows that CNN-CDV and FFNN-SS are as accurate as MLE, while FFNN-CDV has a lower accuracy, in terms of bias, absolute and relative error, for the speciation, extinction, net diversification and turnover rates (**Fig. 2**, **Table S1**). The likelihood of BD model has an exact analytical solution. This implies that MLE, together with CNN-CDV and FFNN-SS, are as accurate as one can be. The good performance of FFNN-SS might be explained by the representation of the lineage-through-time plot in the summary statistics, that contains all information available in the tree for homogeneous BD models (Nee et al. 1994). The lower performance of FFNN-CDV was expected given that FFNN is less adapted to unstructured data. The neural networks seem to avoid cases of high negative bias on the turnover rate compared to MLE: while this bias can reach down to -0.6 with MLE, it never falls behind -0.35 with CNN-CDV and FFNN-SS (**Fig. 2d**). This is probably due to the fact that contrary to MLE, CNN-CDV and FFNN-SS rarely return estimates close to 0 for the turnover rate (**Fig. S1d**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

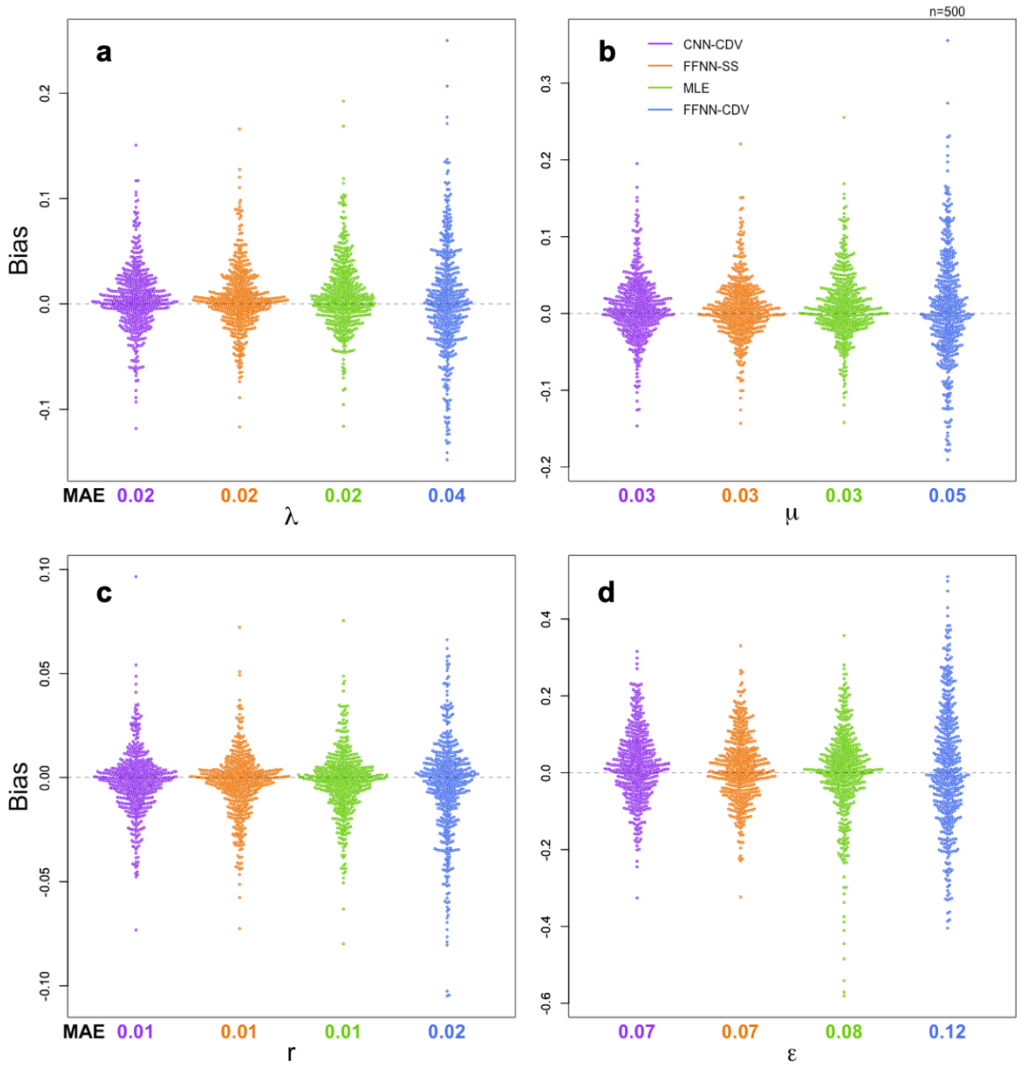

Figure 2: Comparison of estimation accuracy between pretrained CNN-CDV, FFNN-SS, FFNN-CDV and MLE for the BD model.

Swarm plots representing the distribution of estimation biases across 500 simulations. Each dot represents the bias of a single simulation. Estimated parameters were obtained with Convolutional Neural Networks- Complete Diversity-reordered Vector (CNN-CDV in purple), Feed-Forward Neural Networks- Summary Statistics (FFNN-SS in orange), Maximum Likelihood Estimation (MLE in green) and Feed-Forward Neural Networks- Complete Diversity-reordered Vector (FFNN-CDV in blue) for (a) the speciation, (b) the extinction, (c) the net

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

diversification and (d) the turnover rate. The mean absolute error (MAE) is displayed under each swarm plot.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The comparison of parameter estimates obtained using deep learning versus MLE for the BiSSE model confirms that CNN-CDV is at least as accurate as MLE. The only exception is for λ_2 estimates, where MLE implemented in diversitree is slightly more accurate than CNN-CDV in terms of mean absolute error. CNN-CDV is more accurate than the fast MLE algorithm implemented in castor for all parameters (**Fig. 3, Table S2**). The accuracy of the MLE estimation implemented in castor is similar to that of the least reliable FFNN-CDV neural network; it performs slightly better for some parameters (μ_1 and q_{12}) but slightly worse for others (μ_2). The CNN trained on a CDV without the individual tip information (CNN-CDV-less) is much less accurate, indicating that the information on individual tips states is well presented in the CDV representation, and extracted by the CNN (**Table S2**).

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

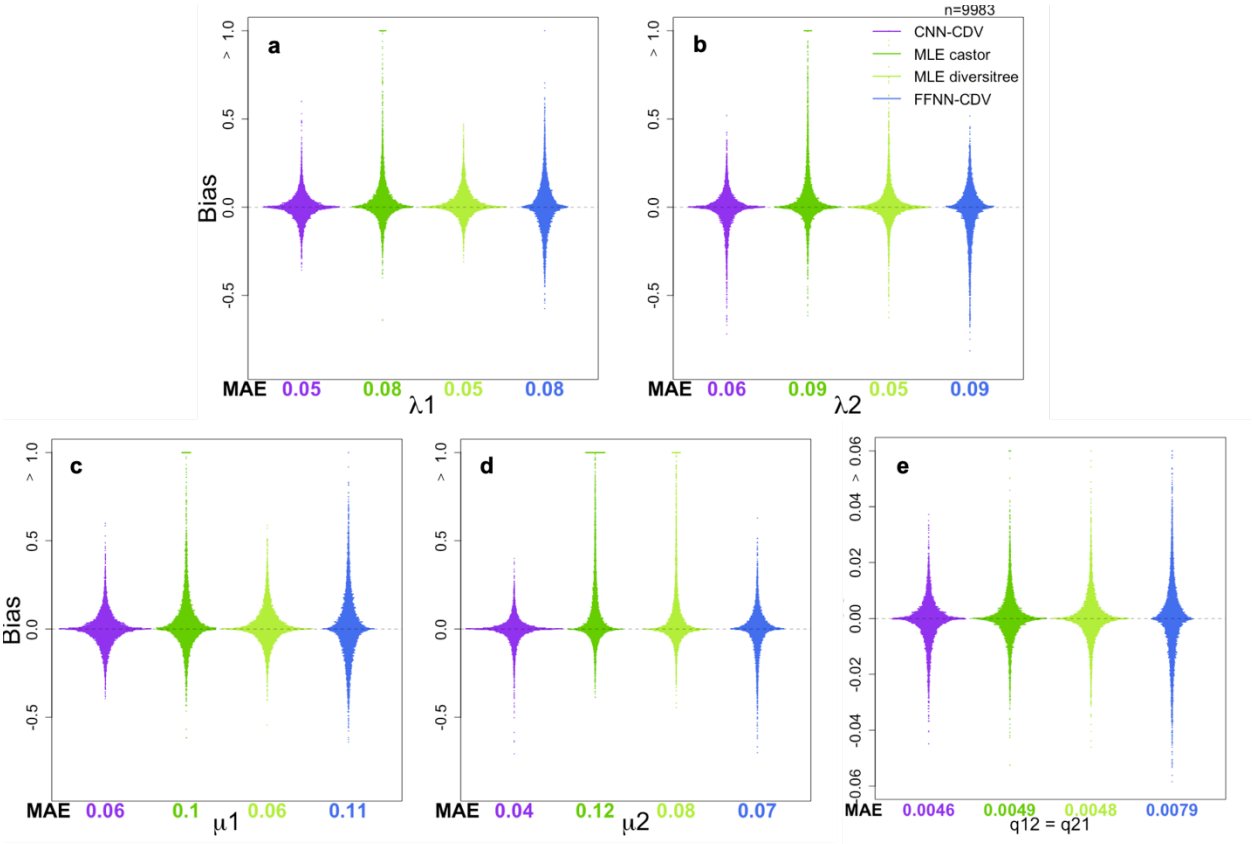

Figure 3: Comparison of estimation accuracy between pretrained CNN-CDV, FFNN-CDV, and MLE obtained with two different inference software for the BiSSE model.

Swarm plots representing the distribution of estimation biases across 9,983 simulations. Estimated parameters were obtained with CNN-CDV (in purple), castor (in dark green), diversitree (in light green) and FFNN-CDV (in blue) for a) the speciation rates 1 and b) 2, c) the extinction rates 1 and d) 2 and e) the transition rates (q_{12}=q_{21}, in our setting). The simulations for which castor or diversitree outputted an error message (15/10,000 for castor and 3/10.000 for diversitree) were excluded from the comparison. The MAE is displayed under each swarm plot. For visualization purposes, the upper outliers of the swarm plot are not displayed at their extreme values but instead are put at the boundaries of the chart.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For both the BD and BiSSE model, the accuracy of CNN-CDV increases as the tree size increases (**Fig. 4**). Similar to MLE, a tree size of at least 300 sampled extant species is required for a median relative absolute error of 6% for the speciation rate and 18% for the extinction rate in the case of the BD model (**Fig. S3**). In the case of the BiSSE model, a tree size of at least 380 is required for a median relative absolute error of less than 14% on the speciation rates, 25% on the extinction rates, and 13% on the transition rate.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

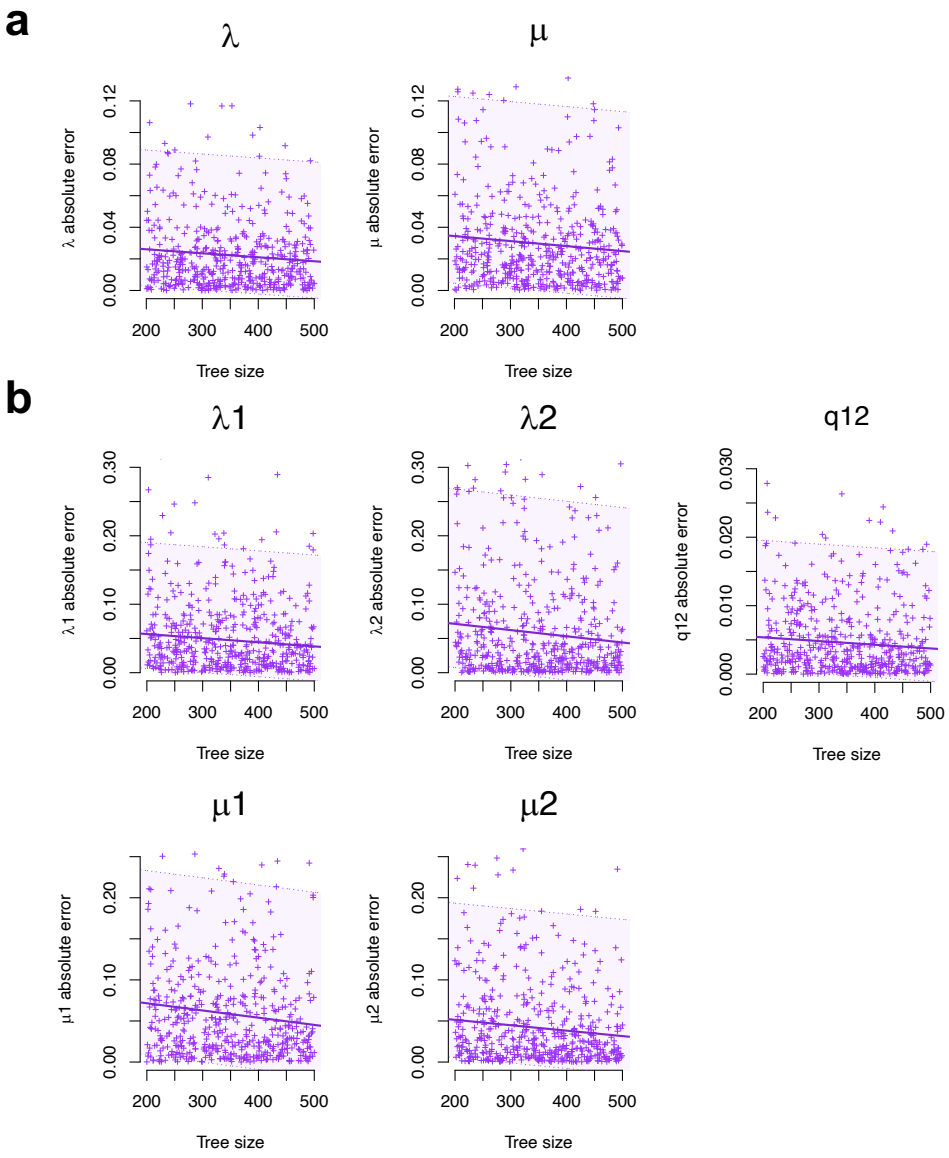

Figure 4: Effect of tree size on the absolute error of parameter estimates when using CNN-CDV under the BD and BiSSE models.

For each model a) constant-rate birth-death (BD), and b) Binary-State Speciation and Extinction (BiSSE), we display the regression on absolute error for each parameter as a function of tree size for 500 test trees (for BiSSE 500 values are shown instead of 10.000 for visualization purposes). The area around the solid line delimited by the dotted lines represent the 95% confidence interval around the regression.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The parameter estimation for the 10.000 BiSSE simulations took 629.5 CPU hours with diversitree, 55.6 CPU hours with castor and 0.4 CPU hours with CNN-CDV, which consisted in encoding the test set into CDV. Once the CNN-CDV is trained, the estimation is thus around 140 times faster than castor and over 1500 times faster than diversitree. Training the network entailed first simulating the training set (1 million trees, 40 CPU hours). The training in itself then took 8 CPU hours. Contrary to MLE for which a large number of CPU hours are required for each new empirical analysis, with deep learning the empirical analyses are very fast once the network has been trained. The same pre-trained networks should be applicable to a large variety of empirical trees, thanks to tree rescaling which enables applications to clades of very different ages and speciation rates.

EMPIRICAL ILLUSTRATION

The primate phylogeny with associated character state (mutualistic or antagonistic interaction with plants) passed the sanity checks on model adequacy. Indeed, all the summary statistics computed on the empirical data fell within the range spanned by the simulations. Likewise, our PCA analysis on these summary statistics showed the empirical data nested in the simulations space, when considering both PC1 and PC2 (explaining together 69% of the variance of our data, **Fig. 5a**) and PC3 and PC4 (that represent an additional 12 points of explained variance, **Fig. 5b**). The CNN-CDV analyses estimated a speciation rate of 0.295 for primate species with a mutualistic interaction with plants, and of 0.093 for those with an antagonistic interaction. The turnover rate ε was estimated at 0.234, and the transition rate at 0.0089. The resulting estimated net diversification rate is 0.225 for primates with a mutualistic interaction, and 0.071 for those with an antagonistic interaction. This simple analysis suggests that mutualistic interactions can

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

favor diversification in primates, as found by (Gómez and Verdú 2012), although we do not interpret this result further here, as models with hidden traits should be used to reach more convincing biological conclusions (Beaulieu and O'Meara 2016). The goal of this empirical analysis is simply to illustrate how the method and trained networks can be used on empirical data.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

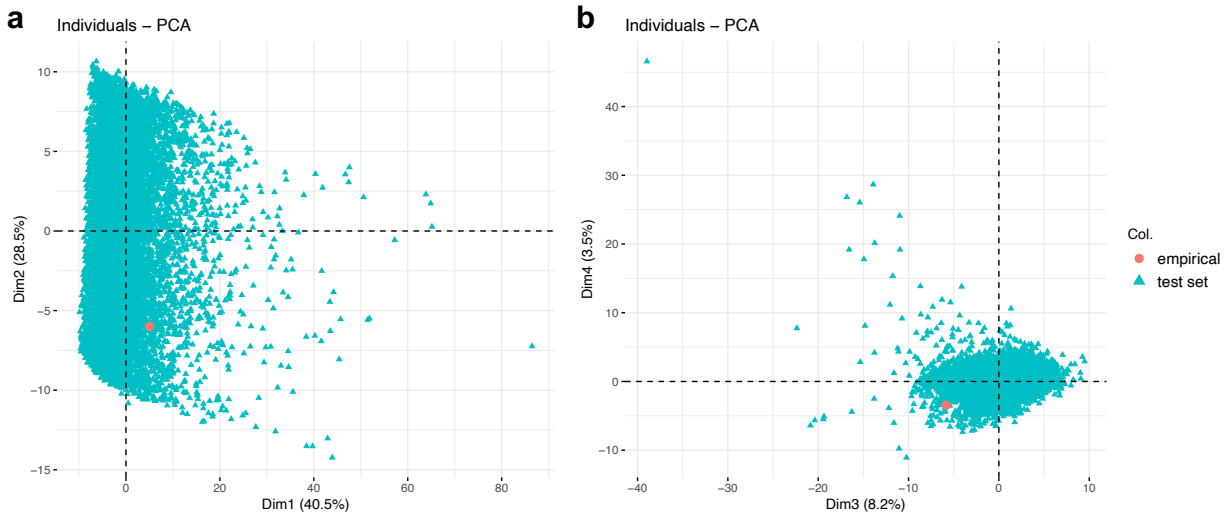

Figure 5: The primate data falls within the space of our BiSSE simulations.

Coordinates of the empirical data (in pink) and of the test set simulations (turquoise) on the a) PC1 and PC2 axes and b) PC3 and PC4 axes of a principal component analysis performed on 102 summary statistics.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DISCUSSION

We developed, tested and illustrated the use of a deep learning based inference approach for phylogenetic diversification analyses, including the case of trait-dependent diversification. We found that both convolutional neural networks combined with a compact representation of the phylogenetic data into a matrix (the CDV) and feed-forward neural networks combined with summary statistics can reach levels of parameter estimation accuracy comparable to those obtained with the well-established maximum likelihood approach, while being faster by several orders of magnitude.

To demonstrate the potential of deep learning for phylogenetic diversification analyses, we worked with two simple diversification models on which MLE estimates can easily be obtained for comparison (the BD and BiSSE models). We also worked with phylogenies of relatively moderate size (200 to 500 extant species sampled). The real value of deep learning will be to allow rapid inferences for more complex models for which likelihoods are not tractable or long to compute and for large phylogenies of several thousands of extant species. Our analyses on simple models demonstrate that it is worth putting efforts and computation power into simulating more complex models, generating larger trees, and training neural networks on such simulations. Given our results, we can expect efficient simulators combined with the CDV representation and CNN to provide an accurate likelihood-free estimation method applicable to very large phylogenies. The CDV representation is not model-specific and can easily be enriched with information on both internal nodes and tips. It could be used for example to represent information on species multidimensional traits, geographic distributions, abundances, and genetic diversity. Combined with efficient simulation models for the evolution of biodiversity (Hagen et al. 2021), the CNN-CDV deep learning inference approach could help adjusting

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

biologically realistic biodiversity models to multifaceted data for a better understanding of how present-day biodiversity was generated, maintained, and distributed geographically.

Parameter estimates are more meaningful when associated with a measure of confidence. With our deep learning framework, obtaining a confidence interval around the estimates can be achieved through approximated parametric bootstrapping (Voznica et al. 2022), which uses the distribution of prediction error measured on simulations. Besides parameter estimation, diversification models are widely used for model comparison, in order to test alternative hypotheses about how diversification proceeds. This problem can be treated with deep learning as a classification problem, with neural networks trained to distinguish simulations from different models. This approach has been developed in Voznica et al. (2022) and shown to perform well. In the case of the BiSSE model for example, a neural network could be trained to distinguish data simulated under a model where states influence speciation rates from data simulated under a model where they do not. Given that speciation rates can be estimated with good accuracy with deep learning, we expect that neural networks will also be able to efficiently distinguish these models given enough differences in speciation rates between states.

We have explored the applicability of CNNs and FFNNs for phylogenetic diversification inference. Other classes of neural networks may also bring accurate solutions for parameter estimation and model selection when combined with simulated datasets. Noticeably, the Graph Neural Networks or Graph Convolutional Networks are neural networks designed for graph data that could be particularly well suited for analyzing phylogenies, encoded as directed graphs. Applications of the GNNs have been developed in the last couple of years across different fields, for instance in protein interaction prediction, drug design or social networks analyses (Zhou et al.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2020). More work is required to assess which network architecture performs best for phylogenetic diversification analyses, which has been initiated in Laajaiti et al..

As the ground truth is unknown, the neural networks are trained on simulations. This raises questions on how robust such approaches are to model misspecification, even though some studies suggest that machine learning approaches might be more robust to model misspecification than other inference approaches (Liang and Jordan 2008; Lee et al. 2010). We illustrated how some sanity checks can be performed to verify that the empirical data falls within the space of simulated data. If it does not, outputs of the neural networks should not be trusted. As in Voznica et al. (2022), we used summary statistics for these sanity checks. Another approach circumventing the use of summary statistics would consist in using autoencoders (Hinton and Salakhutdinov 2006), often deployed for anomaly detection (Chalapathy and Chawla 2019). The autoencoders are a family of NNs, which are trained to output their input while enforcing dimensionality reduction within their neural layers. The induced reconstruction error, *i.e.* the difference between the input and output, can then be used to check if an empirical data is well represented by simulations. For example, autoencoders could be trained on the CDV representation of phylogenetic data simulated under diversification models, and then applied to empirical phylogenetic data. If the reconstruction error is larger for the empirical data than for the simulations, this indicates departures from the simulated model.

We have considered the task of fitting birth-death diversification models to a fixed phylogeny, assumed to be known. While this is a current practice in the field, a better way to account for phylogenetic uncertainty consists in performing full phylogenetic inference, where the phylogenetic tree is inferred from molecular sequence alignments jointly with the parameters of the diversification process (Bouckaert et al. 2014, 2019). Machine learning is already used to

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data mine molecular sequence alignments (Yang et al. 2020), and recent progress has been made for phylogenetic reconstruction as well (Suvorov et al. 2020; Zou et al. 2020; Nesterenko et al. 2022; Solis-Lemus et al. 2022). Ultimately, these recent advances could be in the long term combined to train CNNs directly on sequences simulated from a joint process of speciation, extinction, and sequence evolution.

Deep learning is gaining popularity in biology, including in ecology and evolution (Borowiec et al. 2021). It has been used as a likelihood-free approach to fit population genetic models (Flagel et al. 2019) to sequence data, and more recently to fit epidemiological models to pathogen phylogenies (Voznica et al. 2022). We have shown that it can also perform well as a likelihood-free approach for fitting diversification models to phylogenies of extant species. More work is needed to establish which data representation and network architecture perform best, to perform statistical inference directly on sequence alignments rather than on fixed phylogenies, and to efficiently train networks for more complex models. We hope that our paper will stimulate research in this direction. Ultimately, this should foster the development of new diversification models that are not limited (or whose design is not biased) by our ability to compute likelihoods.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ACKNOWLEDGEMENTS

SL was supported by PSL IRIS Science des données, données de la science and the Fondation pour la Recherche Médicale (FDT202106013269). JV was supported by the Ecole Normale Supérieure Paris-Saclay and by ED Frontières de l'Innovation en Recherche et Education, Programme Bettencourt. HM acknowledges funding from ERC-CoG grant PANDA. We thank Quang Tru Huynh for administrating the GPU farm at Institut Pasteur and the INCEPTION program (Investissement d'Avenir grant ANR16-CONV-0005) that financed the GPU farm.

DATA AVAILABILITY

All data and codes underlying this article are available on GitHub, at https://github.com/JakubVoz/deeptimelearning [doi].

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin M., Ghemawat S., Goodfellow I., Harp A., Irving G., Isard M., Jia Y., Jozefowicz R., Kaiser L., Kudlur M., Levenberg J., Mane D., Monga R., Moore S., Murray D., Olah C., Schuster M., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viegas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y., Zheng X. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. .

Alfaro M.E., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D.L., Carnevale G., Harmon L.J. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proceedings of the National Academy of Sciences. 106:13410–13414.

Andermann T., Antonelli A., Barrett R.L., Silvestro D. 2022. Estimating Alpha, Beta, and Gamma Diversity Through Deep Learning. Frontiers in Plant Science. 13.

Aristide L., Morlon H. 2019. Understanding the effect of competition during evolutionary radiations: an integrated model of phenotypic and species diversification. Ecology Letters. 22:2006–2017.

Avecilla G., Chuong J.N., Li F., Sherlock G., Gresham D., Ram Y. 2022. Neural networks enable efficient and accurate simulation-based inference of evolutionary parameters from adaptation dynamics. PLoS Biol. 20:e3001633.

Barido-Sottani J., Vaughan T.G., Stadler T. 2020. A Multitype Birth–Death Model for Bayesian Inference of Lineage-Specific Birth and Death Rates. Systematic Biology. 69:973–986.

Beaulieu J.M., O'Meara B.C. 2016. Detecting Hidden Diversification Shifts in Models of Trait-Dependent Speciation and Extinction. Syst Biol. 65:583–601.

Beaumont M.A., Cornuet J.-M., Marin J.-M., Robert C.P. 2009. Adaptive approximate Bayesian

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

computation. Biometrika. 96:983–990.

Beaumont M.A., Zhang W., Balding D.J. 2002. Approximate Bayesian Computation in Population Genetics. Genetics. 162:2025–2035.

Bengio Y. 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. In: Montavon G., Orr G.B., Müller K.-R., editors. Neural Networks: Tricks of the Trade: Second Edition. Berlin, Heidelberg: Springer. p. 437–478.

Blum M.G.B., François O. 2010. Non-linear regression models for Approximate Bayesian Computation. Stat Comput. 20:63–73.

Blum M.G.B., Nunes M.A., Prangle D., Sisson S.A. 2013. A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation. Statistical Science. 28:189–208.

Bokma F. 2006. Artificial neural networks can learn to estimate extinction rates from molecular phylogenies. Journal of Theoretical Biology. 243:449–454.

Bokma F. 2010. Time, Species, and Separating Their Effects on Trait Variance in Clades. Systematic Biology. 59:602–607.

Borowiec M.L., Dikow R.B., Frandsen P.B., McKeeken A., Valentini G., White A.E. 2021. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution. 13:1640–1660.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol. 10:e1003537.

Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., Heled J., Jones G., Kühnert D., De Maio N., Matschiner M., Mendes F.K., Müller N.F., Ogilvie H.A., Du Plessis L., Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard M.A., Wu

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

683 C.H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. BEAST 2.5: An advanced software
684 platform for Bayesian evolutionary analysis. PLoS Computational Biology. 15:e1006650–
685 e1006650.
686 Chalapathy R., Chawla S. 2019. Deep Learning for Anomaly Detection: A Survey. .
687 Chollet F.K. 2015. Keras: the Python deep learning API. Available from https://keras.io/.
688 Clevert D.-A., Unterthiner T., Hochreiter S. 2015. Fast and Accurate Deep Network Learning by
689 Exponential Linear Units (ELUs). 4th International Conference on Learning Representations,
690 ICLR 2016 - Conference Track Proceedings.
691 Condamine F.L., Rolland J., Morlon H. 2013. Macroevolutionary perspectives to environmental
692 change. Ecology Letters. 16:72–85.
693 Condamine F.L., Rolland J., Morlon H. 2019. Assessing the causes of diversification
694 slowdowns: temperature-dependent and diversity-dependent models receive equivalent support.
695 Ecology Letters. 22:1900–1912.
696 Cormen T.H. 2009. Introduction to algorithms. Cambridge, Mass: MIT Press.
697 Del Moral P., Doucet A., Jasra A. 2012. An adaptive sequential Monte Carlo method for
698 approximate Bayesian computation. Stat Comput. 22:1009–1020.
699 Dempster A.P., Laird N.M., Rubin D.B. 1977. Maximum Likelihood from Incomplete Data Via
700 the EM Algorithm - Dempster - 1977 - Journal of the Royal Statistical Society: Series B
701 (Methodological) - Wiley Online Library. Available from
702 https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.
703 Etienne R.S., Haegeman B., Stadler T., Aze T., Pearson P.N., Purvis A., Phillimore A.B. 2012.
704 Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record.
705 Proc. R. Soc. B. 279:1300–1309.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

706 Fabre P.-. H., Rodrigues A., Douzery E.J.P. 2009. Patterns of macroevolution among Primates
707 inferred from a supermatrix of mitochondrial and nuclear DNA. Molecular Phylogenetics and
708 Evolution. 53:808–825.
709 FitzJohn R.G. 2010. Quantitative Traits and Diversification. Systematic Biology. 59:619–633.
710 Fitzjohn R.G. 2012. Diversitree: Comparative phylogenetic analyses of diversification in R.
711 Methods in Ecology and Evolution. 3:1084–1092.
712 Flagel L., Brandvain Y., Schrider D.R. 2019. The unreasonable effectiveness of convolutional
713 neural networks in population genetic inference. Molecular Biology and Evolution. 36:220–238.
714 Gamisch A. 2016. Notes on the Statistical Power of the Binary State Speciation and Extinction
715 (BiSSE) Model. Evolutionary Bioinformatics. 12:EBO.S39732.
716 Gillespie D.T. 1977. Exact stochastic simulation of coupled chemical reactions. Journal of
717 Physical Chemistry. 81:2340–2361.
718 Goldberg E.E., Lancaster L.T., Ree R.H. 2011. Phylogenetic Inference of Reciprocal Effects
719 between Geographic Range Evolution and Diversification. Systematic Biology. 60:451–465.
720 Gómez J.M., Verdú M. 2012. Mutualism with Plants Drives Primate Diversification. Systematic
721 Biology. 61:567–577.
722 Goodfellow I., Bengio Y., Courville A. 2016. Deep Learning. .
723 Gubry-Rangin C., Kratsch C., Williams T.A., McHardy A.C., Embley T.M., Prosser J.I.,
724 Macqueen D.J. 2015. Coupling of diversification and pH adaptation during the evolution of
725 terrestrial Thaumarchaeota. Proc Natl Acad Sci USA. 112:9370–9375.
726 Hagen O., Flück B., Fopp F., Cabral J.S., Hartig F., Pontarp M., Rangel T.F., Pellissier L. 2021.
727 gen3sis: the general engine for eco-evolutionary simulations on the origins of biodiversity. .
728 Harmon L.J. 2019. Phylogenetic Comparative Methods - Learning from trees. CC-BY-4.0

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

729 license: .

730 Herrera-Alsina L., van Els P., Etienne R.S. 2019. Detecting the Dependence of Diversification

731 on Multiple Traits from Phylogenetic Trees and Trait Data. Systematic Biology. 68:317–328.

732 Hinchliff C.E., Smith S.A., Allman J.F., Burleigh J.G., Chaudhary R., Coghill L.M., Crandall

733 K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Laughinghouse H.D.,

734 McTavish E.J., Midford P.E., Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams T.,

735 Cranston K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life.

736 Proc Natl Acad Sci USA. 112:12764–12769.

737 Hinton G.E., Salakhutdinov R.R. 2006. Reducing the Dimensionality of Data with Neural

738 Networks. Science.

739 Höhna S. 2014. Likelihood Inference of Non-Constant Diversification Rates with Incomplete

740 Taxon Sampling. PLoS ONE. 9:e84184.

741 Höhna S., Freyman W.A., Nolen Z., Huelsenbeck J.P., May M.R., Moore B.R. 2019. A Bayesian

742 Approach for Estimating Branch-Specific Speciation and Extinction Rates. .

743 Janzen T., Etienne R.S. 2016. Inferring the role of habitat dynamics in driving diversification:

744 evidence for a species pump in Lake Tanganyika cichlids. .

745 Janzen T., Höhna S., Etienne R.S. 2015. Approximate Bayesian Computation of diversification

746 rates from molecular phylogenies: introducing a new efficient summary statistic, the nLTT.

747 Methods in Ecology and Evolution. 6:566–575.

748 Kendall D.G. 1948. On the generalized “birth-and-death” process. Annals of Mathematical

749 Statistics. 19:1–15.

750 Kingma D.P., Ba J.L. 2015. Adam: A method for stochastic optimization. 3rd International

751 Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

752 Kroll A., Engqvist M.K.M., Heckmann D., Lercher M.J. 2021. Deep learning allows genome-
753 scale prediction of Michaelis constants from structural features. PLoS Biol. 19:e3001402.
754 Laudanno G., Haegeman B., Rabosky D.L., Etienne R.S. 2020. Detecting Lineage-Specific
755 Shifts in Diversification: A Proper Likelihood Approach. Systematic Biology.:syaa048.
756 LeCun Y., Bottou L., Bengio Y., Haffner P. 1998. Gradient-based learning applied to document
757 recognition. Proceedings of the IEEE. 86:2278–2323.
758 Lee B.K., Lessler J., Stuart E.A. 2010. Improving propensity score weighting using machine
759 learning. Statistics in Medicine. 29:337–346.
760 Liang P., Jordan M.I. 2008. An asymptotic analysis of generative, discriminative, and
761 pseudolikelihood estimators. Proceedings of the 25th International Conference on Machine
762 Learning.:584–591.
763 Lindsay B.G. 1988. Composite Likelihood Methods. Contemporary Mathematics. 80:221–239.
764 Louca S., Doebeli M. 2018. Efficient comparative phylogenetics on large trees. Bioinformatics.
765 34:1053–1055.
766 Louca S., Pennell M.W. 2020. A General and Efficient Algorithm for the Likelihood of
767 Diversification and Discrete-Trait Evolutionary Models. Systematic Biology. 69:545–556.
768 Louca S., Shih P.M., Pennell M.W., Fischer W.W., Parfrey L.W., Doebeli M. 2018. Bacterial
769 diversification through geological time. Nat Ecol Evol. 2:1458–1467.
770 Maddison W.P., Midford P.E., Otto S.P. 2007. Estimating a Binary Character's Effect on
771 Speciation and Extinction. Systematic Biology. 56:701–710.
772 Maliet O., Hartig F., Morlon H. 2019. A model with many small shifts for estimating species-
773 specific diversification rates. Nature Ecology & Evolution. 3:1086–1092.
774 Maliet O., Morlon H. 2022. Fast and Accurate Estimation of Species-Specific Diversification

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

775 Rates Using Data Augmentation. Systematic Biology. 71:353–366.

776 Marin J.-M., Pudlo P., Robert C.P., Ryder R.J. 2012. Approximate Bayesian computational

777 methods. Stat Comput. 22:1167–1180.

778 May M.R., Höhna S., Moore B.R. 2016. A Bayesian approach for detecting the impact of mass-

779 extinction events on molecular phylogenies when rates of lineage diversification may vary.

780 Methods in Ecology and Evolution. 7:947–959.

781 McKay M.D., Beckman R.J., Conover W.J. 1979. A Comparison of Three Methods for Selecting

782 Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics.

783 21:239–239.

784 McPeek M.A. 2008. The Ecological Dynamics of Clade Diversification and Community

785 Assembly. The American Naturalist. 172:E270–E284.

786 Morlon H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters. 17:508–

787 525.

788 Morlon H., Parsons T.L., Plotkin J.B. 2011. Reconciling molecular phylogenies with the fossil

789 record. Proceedings of the National Academy of Sciences. 108:16327–16332.

790 Nee S., May R.M., Harvey P.H. 1994. The reconstructed evolutionary process. :7.

791 Nesterenko L., Boussau B., Jacob L. 2022. Phyloformer: towards fast and accurate phylogeny

792 estimation with self-attention networks. :2022.06.24.496975.

793 Pedregosa F., Michel V., Grisel OLIVIERGRISEL O., Blondel M., Prettenhofer P., Weiss R.,

794 Vanderplas J., Cournapeau D., Pedregosa F., Varoquaux G., Gramfort A., Thirion B., Grisel O.,

795 Dubourg V., Passos A., Brucher M., Perrot andÉdouardand M., Duchesnay andÉdouard,

796 Duchesnay EDOUARDDUCHESNAY Fré. 2011. Scikit-learn: Machine Learning in Python

797 Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA,

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

798 VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. Journal of Machine Learning Research.
799 12:2825–2830.
800 Prangle D. 2017. Adapting the ABC Distance Function. Bayesian Analysis. 12:289–309.
801 Pyron R.A., Wiens J.J. 2013. Large-scale phylogenetic analyses reveal the causes of high
802 tropical amphibian diversity. Proc. R. Soc. B. 280:20131622.
803 Rabosky D.L. 2014. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-
804 Dependence on Phylogenetic Trees. PLoS ONE. 9:e89543.
805 Rabosky D.L., Chang J., Title P.O., Cowman P.F., Sallan L., Friedman M., Kaschner K., Garilao
806 C., Near T.J., Coll M., Alfaro M.E. 2018. An inverse latitudinal gradient in speciation rate for
807 marine fishes. Nature. 559:392–395.
808 Raynal L. 2019. Bayesian statistical inference for intractable likelihood models. .
809 Richter F., Haegeman B., Etienne R.S., Wit E.C. 2020. Introducing a general class of species
810 diversification models for phylogenetic trees. Statistica Neerlandica. 74:261–274.
811 Rolland J., Condamine F.L., Jiguet F., Morlon H. 2014. Faster Speciation and Reduced
812 Extinction in the Tropics Contribute to the Mammalian Latitudinal Diversity Gradient. PLoS
813 Biol. 12:e1001775.
814 Sanchez T., Cury J., Charpiat G., Jay F. 2020. Deep learning for population size history
815 inference: Design, comparison and combination with approximate Bayesian computation.
816 Molecular Ecology Resources.
817 Saulnier E., Gascuel O., Alizon S. 2017. Inferring epidemiological parameters from phylogenies
818 using regression-ABC: A comparative study. PLoS Comput Biol. 13:e1005416.
819 Sheehan S., Song Y.S. 2016. Deep Learning for Population Genetic Inference. PLoS Comput
820 Biol. 12:e1004845.

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

821 Sisson S.A., Fan Y., Beaumont M. 2018. Handbook of Approximate Bayesian Computation.
822 CRC Press.
823 Sisson S.A., Fan Y., Tanaka M.M. 2007. Sequential Monte Carlo without likelihoods.
824 Proceedings of the National Academy of Sciences. 104:1760–1765.
825 Solis-Lemus C., Yang S., Zepeda-Nunez L. 2022. Accurate Phylogenetic Inference with a
826 Symmetry-preserving Neural Network Model. .
827 Srivastava N., Hinton G., Krizhevsky A., Salakhutdinov R. 2014. Dropout: A Simple Way to
828 Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 15:1929–
829 1958.
830 Stadler T. 2009. On incomplete sampling under birth–death models and connections to the
831 sampling-based coalescent. Journal of Theoretical Biology. 261:58–66.
832 Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of
833 the National Academy of Sciences. 108:6187–6192.
834 Stadler T. 2013. Recovering speciation and extinction dynamics based on phylogenies. Journal
835 of Evolutionary Biology. 26:1203–1219.
836 Stone B.W., Wolfe A.D. 2021. Asynchronous rates of lineage, phenotype, and niche
837 diversification in a continental-scale adaptive radiation. .
838 Suvorov A., Hochuli J., Schrider D.R. 2020. Accurate Inference of Tree Topologies from
839 Multiple Sequence Alignments Using Deep Learning. Systematic Biology. 69:221–233.
840 Varin C., Reid N., Firth D. 2021. AN OVERVIEW OF COMPOSITE LIKELIHOOD
841 METHODS. :39.
842 Vasconcelos T., O'Meara B.C., Beaulieu J.M. 2022. A flexible method for estimating tip
843 diversification rates across a range of speciation and extinction scenarios. Evolution. 76:1420–

bioRxiv preprint doi: https://doi.org/10.1101/2022.09.27.509667; this version posted September 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1433.

Villarreal J., Renner S.S. 2013. Correlates of monoicy and dioicy in hornworts, the apparent sister group to vascular plants. BMC Evol Biol. 13:239.

Voznica J., Zhukova A., Boskova V., Saulnier E., Lemoine F., Moslonka-Lefebvre M., Gascuel O. 2022. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. Nat Commun. 13:3896.

Williams J.H., Taylor M.L., O'Meara B.C. 2014. Repeated evolution of tricellular (and bicellular) pollen. American Journal of Botany. 101:559–571.

Yang A., Zhang W., Wang J., Yang K., Han Y., Zhang L. 2020. Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Frontiers in Bioengineering and Biotechnology. 8:1032–1032.

Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution. 14:717–724.

Zhou J., Cui G., Hu S., Zhang Z., Yang C., Liu Z., Wang L., Li C., Sun M. 2020. Graph neural networks: A review of methods and applications. AI Open. 1:57–81.

Zou Z., Zhang H., Guan Y., Zhang J., Liu L. 2020. Deep Residual Neural Networks Resolve Quartet Molecular Phylogenies. Molecular Biology and Evolution. 37:1495–1507.